Received: 29 January 2022 | Revised: 29 March 2022 | Accepted: 6 April 2022

DOI: 10.1002/pro.4325

FULL-LENGTH PAPER

THE PROTEIN SOCIETY WILEY

What the protein data bank tells us about the evolutionary conservation of protein conformational diversity

Mallika Iyer[1] | **Lukasz Jaroszewski**[2] | **Mayya Sedova**[2] | **Adam Godzik**[2]

[1]Graduate School of Biomedical Sciences, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, California, USA

[2]Biosciences Division, University of California Riverside School of Medicine, Riverside, California, USA

Correspondence
Adam Godzik, Biosciences Division, University of California Riverside School of Medicine, 900 University Ave., Riverside, CA 92521, USA.
Email: adam.godzik@medsch.ucr.edu

Funding information
Bruce D. and Nancy B. Varner Endowment Fund; National Institute of General Medical Sciences, Grant/Award Number: 118187

Review Editor: John Kuriyan

Abstract

Proteins sample a multitude of different conformations by undergoing small- and large-scale conformational changes that are often intrinsic to their functions. Information about these changes is often captured in the Protein Data Bank by the apparently redundant deposition of independent structural solutions of identical proteins. Here, we mine these data to examine the conservation of large-scale conformational changes between homologous proteins. This is important for both practical reasons, such as predicting alternative conformations of a protein by comparative modeling, and conceptual reasons, such as understanding the extent of conservation of different features in evolution. To study this question, we introduce a novel approach to compare conformational changes between proteins by the comparison of their difference distance maps (DDMs). We found that proteins undergoing similar conformational changes have similar DDMs and that this similarity could be quantified by the correlation between the DDMs. By comparing the DDMs of homologous protein pairs, we found that large-scale conformational changes show a high level of conservation across a broad range of sequence identities. This shows that conformational space is usually conserved between homologs, even relatively distant ones.

KEYWORDS

conformational changes, conformational ensembles, difference distance maps, evolutionary conservation

SIGNIFICANCE

Proteins do not exist in a single conformation but undergo conformational changes, and these changes are intrinsic to their functions. Here, we show that large-scale conformational changes are highly conserved between homologous proteins across a broad range of evolutionary distances. Due to this conservation, alternative conformations may be predicted for a given protein based on its homologs, leading to more accurate docking, function prediction, and better overall understanding of protein function.

1 | INTRODUCTION

The sequence-structure-function paradigm is a fundamental part of molecular evolutionary biology. We study similarities between proteins' sequences and structures and use them to reason about their evolutionary relations and the similarity of their functions. This paradigm has been extended and reinterpreted many times and an important, outstanding question for its practical applications is which specific features of sequence, structure, and function are conserved between homologs. For instance, it has long been observed that homologous proteins share similar folds, but the level of similarity tends to wane with increasing evolutionary distance and diminishing sequence similarity between the homologs.[1] But other features, such as the stoichiometry of complexes formed by homologs, are less conserved.[2] In this manuscript, we explore the conservation of large-scale conformational changes of proteins to understand if and how they may be predicted based on homology. In their native state, proteins are highly flexible and exist in a multitude of conformations forming an ensemble.[3,4] A protein can occupy different conformations in the ensemble by undergoing conformational movements with a broad range of time and length scales.[5] This flexibility is often intrinsic to protein function[6–8] and thus, in order to understand a protein's function, it is essential to know the conformational changes it undergoes.

There are many experimental methods for studying protein flexibility, such as NMR (nuclear magnetic resonance) relaxation-dispersion experiments[9] and time-resolved crystallography.[10] Computational methods, like all-atom molecular dynamics (MD) simulations[11] and Normal Mode Analysis (NMA) (most often using Elastic Network Models [ENMs][12–15]), can be used to predict the flexibility patterns/alternative conformations of a protein. Homology-based prediction of large-scale conformational changes could provide a simpler alternative, but it would require these changes to be conserved between homologs. Many studies have shown that homologous proteins share similar patterns of structural flexibility, typically by indirect experimental and computational approaches such as normal modes, B-factor profiles, or the NMR relaxation-dispersion constants of various residues.[16–22] However, this was mostly focused on local flexibility involving small-scale conformational changes. Reliable predictions of large-scale conformational changes would be important not only for our general knowledge about a protein's conformational space, but also for many practical applications such as modeling for molecular replacement or for cryoEM or in docking studies. It would also enable the prediction of alternative conformations of a protein based on those on its homologs—an application that has been explored in the ConTemplate[23] and ModFlex[24] servers. Therefore, in this manuscript, we evaluate the conservation of large-scale conformational changes directly using experimentally solved structures deposited in the Protein Data Bank (PDB).[25]

The PDB contains, on average, more than six coordinate sets per individual protein that provide a sample of the protein's conformational ensemble,[26–28] often capturing distinct conformational and functional states of the protein and thus characterizing different "neighborhoods" in the ensemble. A previous study used this multiplicity of coordinate sets to show that protein pairs that share one similar conformation often share multiple conformations—suggesting that their conformational spaces are conserved.[23] Here, we expand on this analysis, using a different approach—instead of directly comparing various conformations of two proteins, we compare their conformational changes. This requires comparing the differences between pairs of conformations (Figure 1). The advantage of this approach is that it would capture the similarity in the conformational changes between proteins that have some distinct structural features (a conceptual example is presented in Figure 1).

Our group previously developed the PDBFlex server[31] to study the flexibility and conformational diversity of proteins using experimentally solved X-ray crystallographic coordinate sets from the PDB. Here, we use the PDBFlex server to further study the similarity of conformational changes in homologous proteins. However, there is no established, systematic method to compare conformational changes.[32] Thus, we first developed a method to do this using the distance map representation of protein structures (Figure 1). For each protein with two distinct conformations, we calculated the difference distance map (DDM) representing the conformational difference between them. The DDMs of pairs of homologous proteins were then compared and the DDM similarities were quantified by calculating the correlation between them. We found that large-scale conformational changes are highly conserved between homologous proteins across a wide range of evolutionary distances, as most homologs had high DDM correlations. This suggests that such conformational changes can be inferred for a given target protein based on the conformational changes of its homologs.

2 | RESULTS

2.1 | Characterization of conformational diversity using X-ray crystallographic coordinate sets from the Protein Data Bank

The PDBFlex server[31] identifies groups of independently solved coordinate sets of the same protein, which we call "clusters." We divided each PDBFlex cluster into subclusters representing distinct conformations of the protein (or neighborhoods in the ensemble) based on a 3 Å RMSD

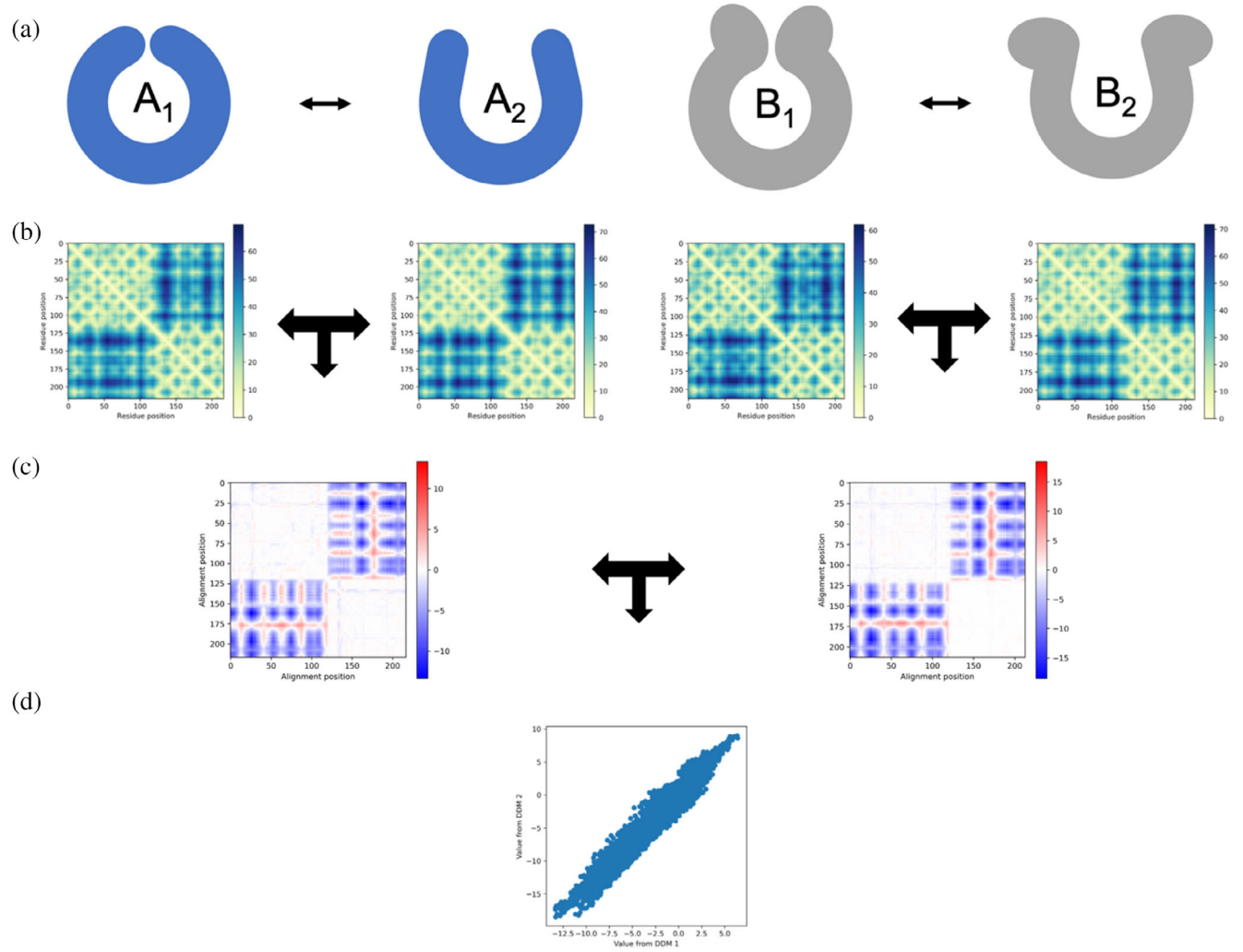

FIGURE 1 Analysis of protein conformational changes using difference distance maps (DDMs) (a) Two proteins (A and B) with significant structural differences have two conformations each (A_1 and A_2, B_1 and B_2) and undergo similar conformational changes, such that the difference between the conformations of A (A_2-A_1) is similar to the difference between the conformations of B (B_2-B_1). (b) The conformations are described by distance maps. (c) Differences between conformations are described by DDMs. (d) Similarities between DDMs can be measured by their correlation. PDB chains used to make the DMs and DDMs: mouse catalytic antibody 39-A11 1a4kH[29] and 1a4jB,[29] and *Llama glama* Fab 48A2 anti-Met antibody 4r96B[30] and 4r96F[30]

threshold. This allowed us to focus on large-scale conformational changes such as relative domain rearrangements.[27] One representative coordinate set was selected for each subcluster to use for further analyses (see Methods). We found that with the 3 Å threshold, most proteins (~93%) have only one distinct conformation represented in the PDB, but there are over 2,000 proteins for which there are at least two conformations (Figure S1).

2.2 | Identification of homologous protein pairs and distribution of protein families

We next identified homologous protein pairs to compare their conformational changes. For simplicity, we only considered proteins with exactly two conformations in our dataset since this would limit the comparison to just two pairs of coordinate sets (four coordinate sets in total) per homologous pair (Figure 1). Briefly, from the set of proteins with two conformations, a total of 48,489 homologous pairs were identified using BLAST.[33] To ensure that the comparison of conformational changes would be based on the full length of both proteins, these pairs were further filtered such that both the query and the subject sequence had $\geq$90% coverage in the alignment (see Methods). This resulted in a final set of 530 proteins forming 20,740 pairs.

We then assessed the distribution of protein families in this dataset, by mapping each protein to its corresponding Pfam[34] family(ies) using HMMER.[35] Surprisingly, the final set of homologous pairs had a

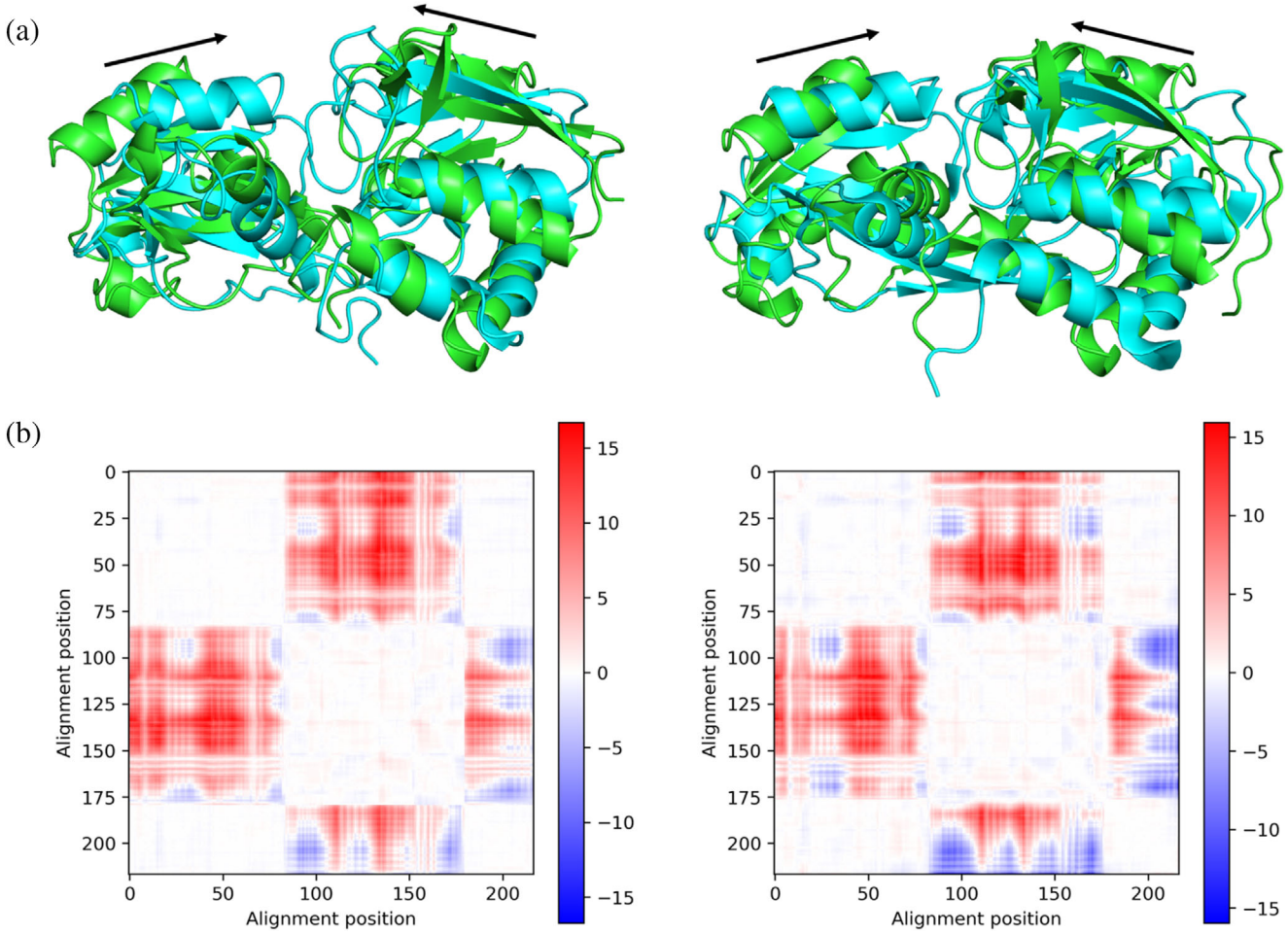

FIGURE 2 Periplasmic binding proteins (PBPs) undergo similar conformational changes and have visually similar difference distance maps (DDMs). (a) Two conformations of: Left: Lysine-Arginine-Ornithine binding protein represented by 2laoA[38] in green and 1lahE[39] in cyan; Right: GlnP substrate-binding domain 2 (SBD2) represented by 4kr5B[37] in green and 4kqpA[37] in cyan (ligands not shown here). (b) DDM of: Left: 2laoA-1lahE; Right: 4kr5B-4kqpA

total of 20,185 (97%) pairs in which either one or both homologs were mapped to the immunoglobulin superfamily/clan. However, only 228 (~13%) of the 1,815 proteins with two conformations were mapped to this superfamily. The overrepresentation of this superfamily in the final set of pairs could be explained largely by the high level of similarity between members of this superfamily (Figure S2). Indeed, the average number of blast hits per query protein from this superfamily was 195.4 (before filtering for coverage), whereas proteins not in this superfamily only had an average of 5.1 hits. To prevent our final conclusions from being biased by the overrepresentation of the immunoglobulin superfamily in our dataset, the homologous pairs were divided into two subsets—immunoglobulin pairs (20,185) and non-immunoglobulin pairs (555)—which were analyzed separately.

2.3 | Representing large-scale conformational changes of proteins using difference distance maps

We next developed a method to systematically compare conformational changes between proteins, based on the distance map representation of protein structures. A distance map (DM) is a matrix of the inter-residue distances of all residue pairs in a protein and offers an alternative representation of protein structures. A protein that undergoes a conformational change can, therefore, be described by two DMs (one for each conformation). The difference between the two conformations (that is, the conformational change) can be represented by a difference distance map (DDM), obtained by subtracting one DM from the other. We calculated DDMs between the representatives of the conformational subclusters for all the proteins in our dataset (Figure 1). A visual

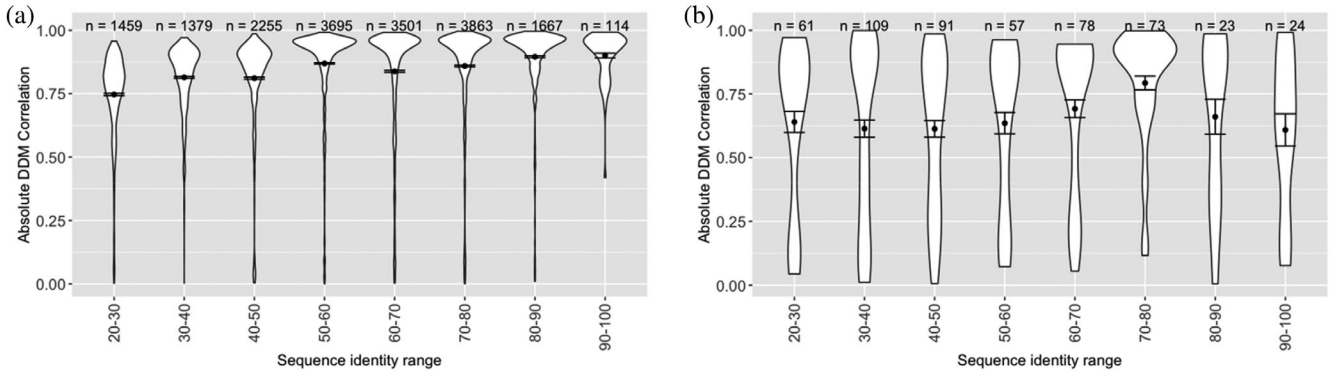

FIGURE 3 Absolute Pearson DDM correlation vs. sequence identity for (a) immunoglobulin homologs, p-value $= 1.54 \times 10^{-102}$. (b) non-immunoglobulin homologs, p-value $= 0.0259$. 'n' represents the number of pairs in each bin, points represent means, and error bars represent standard error of the mean. p-values are based on a linear regression of absolute DDM correlation vs. sequence identity, as implemented in R v.4.0.0

comparison of the DDMs and "morphing movies" for several protein pairs suggested that proteins undergoing conformational changes that look similar on visual inspection often have visually similar DDMs. For example, periplasmic binding proteins undergo a typical, "Pacman-like," "close-open" hinge movement upon binding/releasing their substrates[36,37] and have strikingly similar DDMs (Figure 2).

The correlation between the values of equivalent elements of the two DDM matrices offers a simple metric of their similarity. For each pair of homologous proteins in our final dataset, we calculated both the Pearson and Spearman correlation between their DDMs. Both coefficients were well correlated with each other, with the Spearman correlation generally having a lower value (Figure S3). For example, the visual similarity of the DDMs in Figure 2 is reflected in the high values of the DDM correlations which are 0.88 for the Pearson correlation and 0.72 for the Spearman correlation. In the following analyses, we use the absolute value of the correlation to quantify the similarity between two DDMs, as the sign of the correlation simply reflects the arbitrary order in which the individual DMs were subtracted to get the DDM.

2.4 | Conformational changes are highly conserved across a wide range of evolutionary distances

It has long been observed that the structural similarity of homologous proteins decreases with their broadly defined evolutionary distance, usually estimated by the proxy of sequence identity.[1] We explored the extension of this observation to the conformational changes of proteins by analyzing the DDM correlations of homologous

TABLE 1 Distribution of absolute Pearson DDM correlation values for homologous pairs

All homologs	Abs. Pearson correlation ≥ 0.50	Total
Immunoglobulins	16,813 (93.8%)	17,933
Non-immunoglobulins	365 (70.7%)	516
Distant homologs	**Abs. Pearson correlation ≥ 0.50**	**Total**
Immunoglobulins	4,775 (93.8%)	5,093
Non-immunoglobulins	172 (65.9%)	261

protein pairs and found that, for a broad range of sequence similarity, the majority of proteins show high DDM correlations (absolute Pearson correlation ≥ 0.50 or absolute Spearman correlation ≥ 0.30, values based on data shown in Figure S3), suggesting highly similar, that is, conserved pattern of conformational changes (Figure 3, Figure S4, Table 1, Table S1). This was found to be the case even between distant homologs (defined here as having sequence identity <50%). High DDM correlations were observed for more than 90% of the distant immunoglobulin homologs and more than 60% of the distant non-immunoglobulin homologs (Table 1 and Table S1). This confirms the validity of the main hypothesis evaluated here, of the broad conservation of large-scale conformational changes in homologous proteins. An example of such similarity in a pair of distant homologs is illustrated and discussed in detail below for two adenylate-forming enzymes (Figure 4).

The structures of enzymes from the adenylate-forming superfamily consist of an N- and C-terminal domain. These enzymes catalyze two-step reactions and

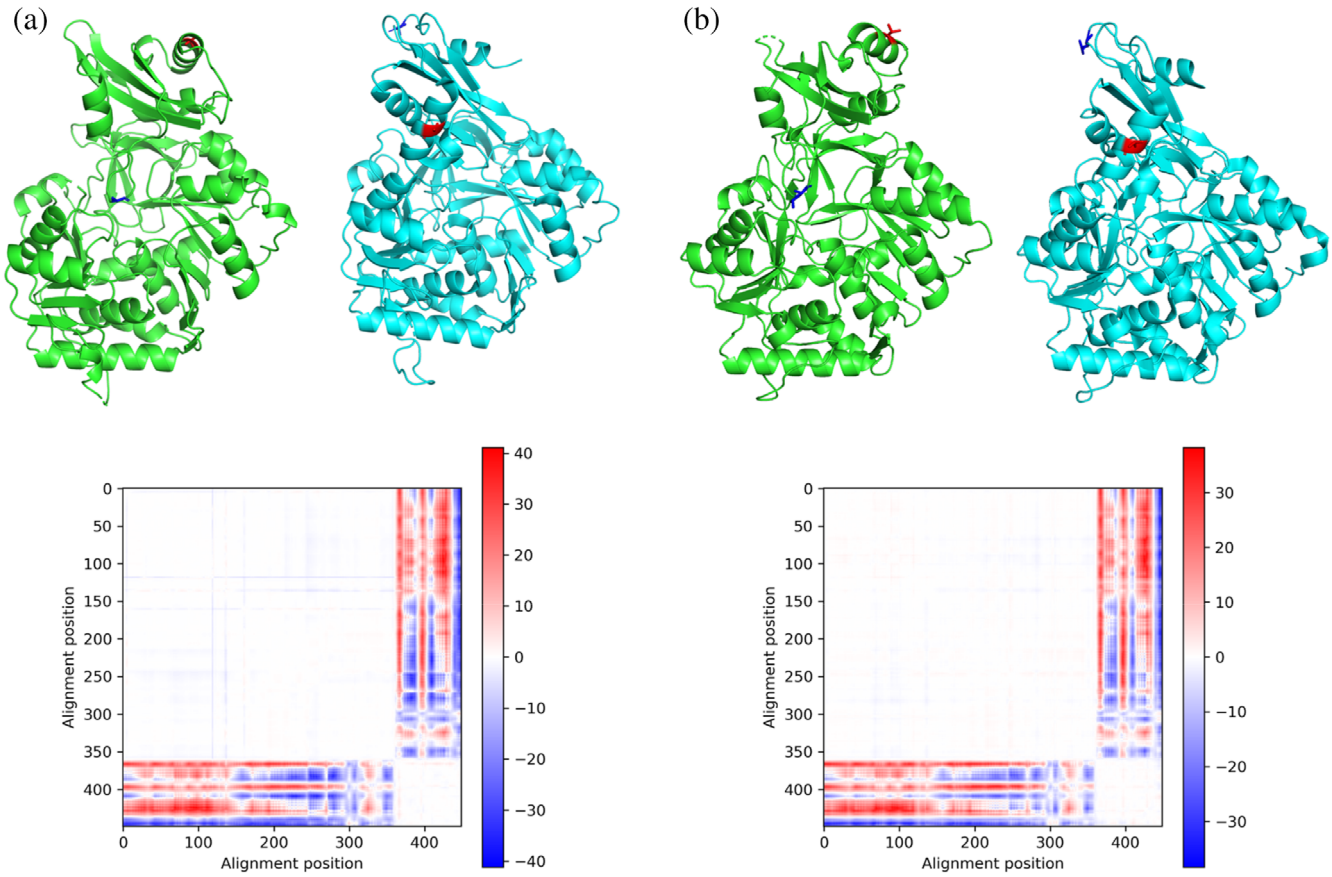

FIGURE 4 Conformational change in two adenylate-forming enzymes. (a) Top: Adenylation conformation (6sq8C[40] in green) and amidation conformation (6sq8E[40] in cyan) of McbA (ligands not shown). Ala455 is shown in red sticks and Ala481 is shown in blue sticks on both conformations for reference. Bottom: The DDM of 6sq8C-6sq8E. (b) Top: Adenylation conformation (3cw8X[41] in green) and thioester-forming conformation (3cw9B[41] in cyan) of 4-chlorobenzoate:CoA ligase (ligands not shown). Thr463 is shown in red sticks and Leu490 is shown in blue sticks on both conformations for reference. Bottom: The DDM of 3cw8X-3cw9B

occupy two different conformations for the catalysis of each half-reaction. The conformational movement involves a large-scale rotation of the C-terminal domain with respect to the N-terminal domain, such that two different faces of the C-terminal domain are presented to the active site for each half-reaction.[42] Our dataset contains one pair of distant homologs from this superfamily, with a sequence identity of 28.36%. The first protein is McbA, a fatty acid CoA ligase from *Marinactinospora thermotolerans*. This enzyme catalyzes the synthesis of β-carboline amides from 1-acetyl-3-carboxy-β-carboline[40,43] by first adenylating the substrate, followed by amidation to give the product.[40] The second protein is 4-chlorobenzoate:CoA ligase from *Alcaligenes sp*. This enzyme catalyzes the adenylation of 4-chlorobenzoate (4-CB), followed by thioesterification to give 4-chlorobenzoate:CoA (4-CB-CoA).[41] Both enzymes occupy two conformations (Figure 4) that reflect the large-scale rotation of the C-terminal domain that is unique to this superfamily. The extreme similarity in the domain rotation in both enzymes is clearly reflected in the similarity of the DDMs which have a Pearson correlation of 0.97 (Spearman correlation of 0.70) (Figure 4). This high similarity is seen despite the low sequence identity, suggesting that if only one of the two conformations had been solved for either of these proteins, the second conformation, and thus, the conformational change, could have been modeled based on that of its homolog.

When the DDM correlations between homologous proteins were compared against their sequence identity, we found that DDM correlation decreases, however slightly, with decreasing sequence identity (Figure 3, Figure S4). This trend is clearly visible for the set of homologous immunoglobulin pairs, but much weaker for non-immunoglobulins (Figure 3 and Figure S4). This observation suggests that the similarity of the conformational changes of a pair of homologs depends, at least to some degree, on their evolutionary distance.

While most homologous pairs in the dataset showed high DDM correlation values, we note that in each

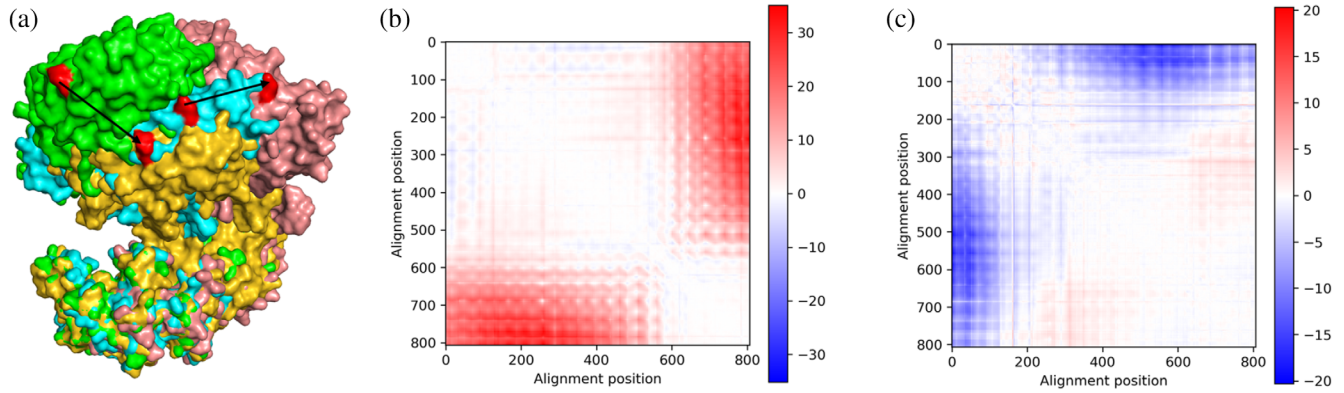

FIGURE 5 Conformations of importin-β homologs. (a) 3ea5D[45] in green, 5owuA[46] in gold, 4xriA[47] in cyan, and 4xrkA[47] in pink. Glu770 in 3ea5D and 5owuA and Asn780 in 4xriA and 4xrkA are shown in red and are connected by arrows. This figure was created by superimposing residues 1–150 of 3ea5D, 5–149 of 4xriA, and 38–149 of 4xrkA onto residues 1–150 of 5owuA based on the sequence alignment. (b) DDM of 3ea5D-5owuA. (c) DDM of 4xriA-4xrkA

sequence identity bin, there is a tail of outliers with low correlations. Manual examination showed that most of these outliers do, in fact, show very different conformational changes. For example, a pair of homologous importin-β proteins from *Saccharomyces cerevisiae* and *Chaetomium thermophilum* (sequence identity ~40%) illustrates this effect (Figure 5). Proteins in the importin-β (Impβ) superfamily are transport proteins that move cargo from the cytoplasm into the nucleoplasm through the nuclear pore complex.[44] In our dataset, there are two conformations for both the *S. cerevisiae* Impβ (represented by 3ea5D[45] and 5owuA[46]) and *C. thermophilum* Impβ (represented by 4xriA and 4xrkA[47]) (Figure 5a). The DDMs for these proteins are strikingly different with a Pearson correlation of 0.023 (Spearman correlation of 0.067) (Figures 5b and C). This is confirmed by visual inspection of the four coordinate sets, which reflect four significantly different conformations. Both homologs in this dataset have one "extended" conformation and one "compressed" conformation. However, the relative directions of the conformational change between the two conformations are different in the two proteins (Figure 5a).

In most cases, we do not know if the lack of correlation between the DDMs of two proteins is caused by real differences between their conformational ensembles or by inadequate sampling of these ensembles in the PDB. This example seems to belong to the latter class, as further analysis showed that 3ea5D represents importin-β bound to RanGTP, while 5owuA is bound to the C-terminal region of the nucleoporin Nup1p. On the other hand, 4xriA and 4xrkA represent the unbound protein in different cellular environments (the polar cytoplasm or nucleoplasm for 4xriA and the apolar nuclear pore channel for 4xrkA). Therefore, the conformations seen here represent two different causes of conformational changes—between binding of two different partners vs. changes in the environment for the apo-structure.

3 | DISCUSSION

Proteins are highly flexible and sample multiple conformations in their conformational ensembles. In this manuscript, we examined the similarities of the large-scale conformational changes of homologous proteins, as sampled by the multiple depositions in the Protein Data Bank. Previous studies have suggested that the flexibility patterns and conformational space of proteins are conserved.[16–23] Here, we expand the analysis of the conservation of large-scale conformational changes to a set of homologous proteins using experimentally solved structures and a newly developed method based on difference distance map (DDM) correlations (see Figure 1 for a visual illustration). The main advantage of the method presented here is that it makes use of experimentally solved structures, thus avoiding assumptions made in computational methods like normal mode analysis (NMA)[12,13] and molecular dynamics (MD).[11] Difference distance maps (DDMs) further offer easy visualization of the structural differences, highly complementary to the usual structure superpositions. We leveraged the multiplicity of coordinate sets in the Protein Data Bank (PDB),[25] as captured by the PDBFlex server,[31] to identify a set of 1,815 proteins with two well-separated conformations. This was done using a 3 Å RMSD threshold. When the threshold is set to lower values, a greater number of distinct conformations can be identified for a given protein. However, this would focus on local/small-scale conformational changes, whereas the goal of this manuscript was to analyze large-scale conformational changes, like

domain rearrangements. The analysis can obviously be repeated with lower thresholds, and we are planning to release a server where users can set up their own thresholds and repeat the analyses. Having identified proteins with two conformations, we then identified homologous proteins pairs and compared their conformational changes based on their DDM correlations. We found that, on average, when conformational ensembles contain two main conformations, the conformational change between them is very similar for homologous proteins. Importantly, this was observed even for very distant homologs (sequence identities in the range of 20–30%) which could mean that large-scale conformational changes are conserved even if precise biochemical functions are not.

The results presented here illustrate both the strength and weakness of using experimentally solved structures to characterize the conformational ensembles of proteins. The PDB depositions provide only a sample of the conformational ensemble of any given protein. Thus, for proteins that sample many different conformations, the PDB may not contain coordinate sets corresponding to all functionally relevant ones. This is evident in a number of outliers with unusually low DDM correlations. Many of these outliers represent homologous proteins that are solved in different conformations, which may simply reflect an incomplete sampling of their ensembles. This observation leads to a practical application, where one could create models of "missing" conformations for individual proteins and ask whether they exist in nature. In the case of the importin-β homologs shown in Figure 5, this is likely to be the case, as the four coordinate sets represent different environmental conditions and/or binding partners of the proteins.[45–47] However, the sampling of the conformational space for most proteins is sufficient to strongly support the general trend of conservation of large-scale conformational changes in homologous proteins.

Besides evidence of the broad conservation of conformational changes, we also observed a slight trend of increasing DDM correlation with increasing sequence identity. Since sequence identity is a widely used (albeit poor) proxy for evolutionary distance, these results suggest that the similarity in conformational changes, like folds,[1] is dependent on evolutionary distance and decreases with increasing distance. A similar observation has also been made for the backbone flexibility profiles of homologous proteins, as characterized by their B-factor profiles.[19] However, the correlation between DDM correlation and sequence identity was particularly strong in the immunoglobulin superfamily and much weaker for the remaining set of proteins. This could be because the sampling of the immunoglobulin family is particularly dense and because many members of this family have similar functions. The remaining proteins, representing a variety of different protein families with different folds, may have more complicated conformational spaces with more potential conformations that are unevenly sampled in the PDB. Further studies looking deeper into individual protein families could help to confirm this.

Overall, the conservation of large-scale conformational changes shown in this study suggests that homology-based modeling of individual conformations of a protein can be extended to multiple conformations. This application was originally explored in the ConTemplate server,[23] which is currently unavailable. We recently developed the ModFlex server[24] as another tool for this purpose. By providing multiple template structures from each homolog identified for a query protein, the user can explore and model a variety of different conformations for the target.

The results shown here also point to a relatively simple method to model/predict the conformational movement of a given target protein. If two different conformations of the target protein can be modeled, the conformational movement between them can be simulated/modeled using a variety of methods. These range from simple morphing algorithms[48,49] to more complex steered molecular dynamics simulations[50] and motion-planning techniques.[51] This was demonstrated for the pore domain of the *Streptomyces lividans* K-channel (KcsA).[51] This kind of modeling has wide applicability to the field of biology. For example, it would make it possible for biologists to analyze the role of specific residues in enabling conformational movements, to perform in silico docking to different conformations including intermediate states, and in general, to form a more complete picture of protein function.

4 | MATERIALS AND METHODS

4.1 | PDBFlex dataset

This project leveraged data from the PDBFlex server.[31] Briefly, the PDBFlex server clusters all X-ray crystallographic coordinate sets from the Protein Data Bank (PDB)[25] using a 95% sequence identity threshold, creating clusters of coordinate sets corresponding to individual proteins in the PDB (while allowing for a few mutations between individual coordinate sets). Each such cluster is represented by one coordinate set (referred to as the cluster master/representative). For each cluster, pairwise CαRMSDs (root mean square deviations of the Cα atoms after optimal superposition, based on their sequence alignments) were calculated between all cluster members and stored as an all-to-all RMSD matrix.

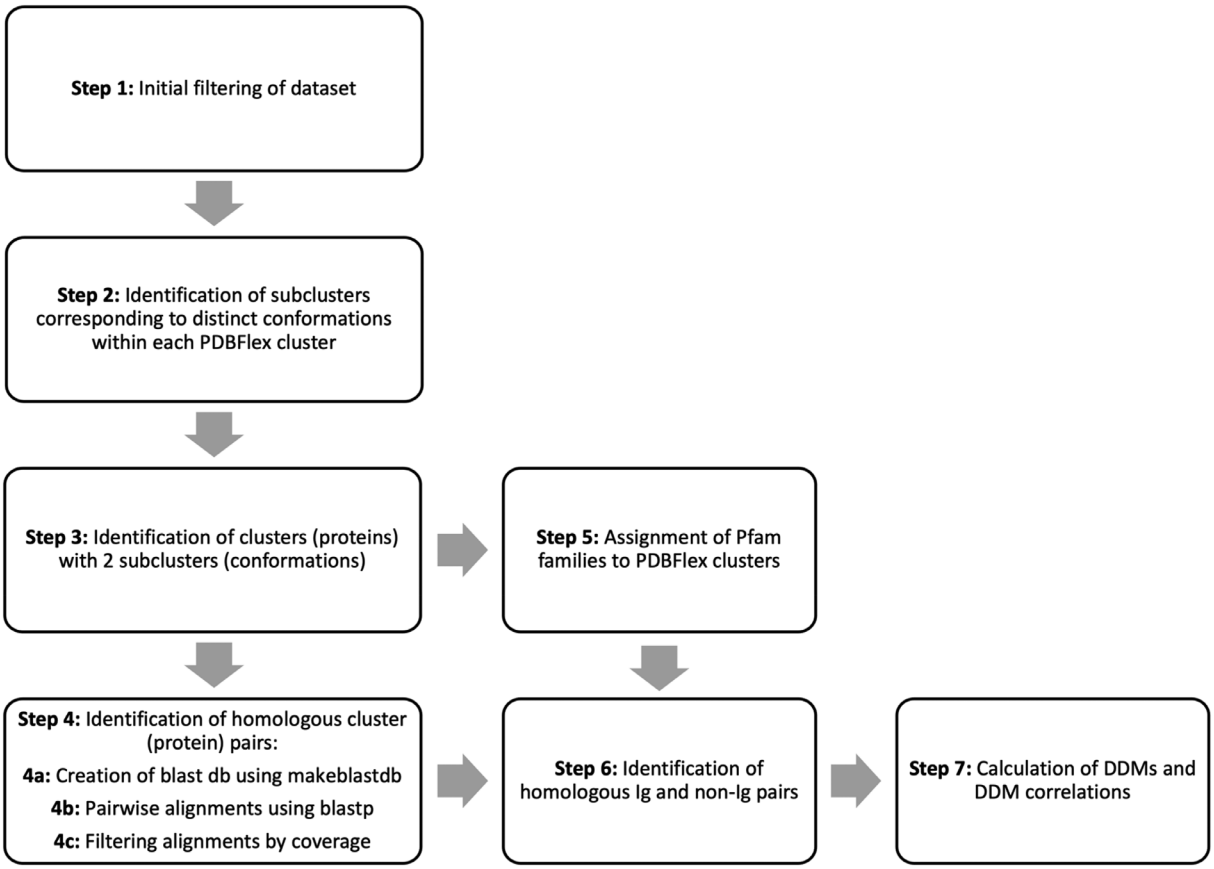

FIGURE 6 Overview of all analyses. Ig, immunoglobulin

The PDBFlex server is automatically updated approximately once per month. All analyses in this project were performed using the November 2, 2020 version of PDBFlex. These analyses/steps are described below, and an overview is presented in Figure 6.

4.2 | Initial filtering of dataset

The version of PDBFlex used for this manuscript contained 364,133 coordinate sets in total. This dataset was filtered (Figure 6, step 1) based on a comparison of the SEQRES sequence (the sequence of the construct used for crystallization) and the PDB sequence (that is, the sequence of residues that were resolved in the structure), using in-house scripts. For each coordinate set:

1. The "maximum possible SEQRES coverage" was calculated as:

$$m = \frac{\text{Length of PDB sequence}}{\text{Length of SEQRES sequence}} \times 100$$

2. The SEQRES sequence and PDB sequence were aligned using an in-house script that uses BLAST.[33] The sequence identity of this alignment was calculated as:

$$s = \frac{\text{No. of identical residues}}{\text{No. of mutually aligned residues(i.e. residues not aligned to gaps)}} \times 100.$$

Coordinate sets with $m < 90\%$ and/or $s < 100\%$ were removed from the dataset. Any coordinate sets for which these values could not be calculated were also removed to ensure that the final dataset did not contain any coordinate sets that did not meet the filtering criteria.

4.3 | Identification of subclusters corresponding to distinct conformations within each PDBFlex cluster

For each PDBFlex cluster with more than one coordinate set, the coordinate sets were grouped based on the CαRMSD matrix into "subclusters" corresponding to distinct conformations (Figure 6, step 2). This grouping was done using a greedy clustering algorithm, as described by Daura *et al.*[52] Additionally, one coordinate set was chosen as the representative of each subcluster, to be used in further analyses. The procedure is described below:

The algorithm first identifies "neighbors" for each coordinate set in a cluster. Two coordinate sets are considered to be neighbors if the RMSD between them is below a predefined threshold. The coordinate set with the maximum number of neighbors is then selected as the representative of the first subcluster, which is composed of this representative and its neighbors. These coordinate sets are then removed from the cluster and the process is repeated until all coordinate sets have been grouped into subclusters.

We set the RMSD threshold to 3 Å in order to analyze large-scale conformational changes. This threshold has been historically used in the field of structure prediction to distinguish correct and incorrect models, and in our earlier analysis,[27] it clearly identified large conformational changes from local ones.

4.4 | Identification of homologous protein pairs

BLAST (v.2.2.30+)[33] was used to identify homologous protein (cluster) pairs. Only proteins (clusters) with two conformations (subclusters) were considered (Figure 6, steps 3–4). First, a FASTA file containing the SEQRES sequence of each cluster master/representative was created (1,815 sequences in total) (Figure 6, step 3). This was used to create a blast database, using makeblastdb (with the -hash_index option) (Figure 6, step 4a).

Pairwise sequence alignments were then obtained by running blastp (Figure 6, step 4b). The FASTA of master sequences was used as the query and the database created in the previous step was used as the search database. The e-value threshold was set to 0.005 and -max_target_seqs to 1815. The output from this step was then filtered such that only the alignments in which the coverage of both the query and subject sequence was $\geq 90\%$ were retained (Figure 6, step 4c). Query coverage was defined as:

$$\frac{\text{No. of query residues in alignment} - \text{No. of query residues aligned to gap}}{\text{Query length}} \times 100$$

where,

No. of query residues in alignment = End of alignment in query – Start of alignment in query +1

No. of query residues aligned to gap = Alignment length – (End of alignment in subject – Start of alignment in subject +1)

(Subject coverage was defined in the same way, except values for query and subject in the above formula were switched.)

In this way, a total of 20,740 homologous protein pairs (i.e., pairs of PDBFlex clusters) were identified where each protein had two conformations (i.e., each PDBFlex cluster contained exactly two subclusters).

4.5 | Assignment of Pfam families to PDBFlex clusters

For each cluster (protein) with two subclusters (distinct conformations), the cluster master/representative sequence was used to identify the corresponding Pfam[34] families (Figure 6, step 5). This was done by running hmmscan (HMMER v.3.3.2)[35] against the Pfam database (Pfam-A, v.34.0) with the –tblout, —dombtblout and –cut_ga options.

4.5.1 | Immunoglobulin superfamily

Clusters that mapped to the immunoglobulin clan/superfamily were identified by parsing the "tblout" file and retrieving all queries (i.e., cluster representatives) that had a hit to at least one of the following Pfam families: Adeno_E3_CR1, Adhes-Ig_like, bCoV_NS7A, bCoV_NS8, C1-set, C2-set, C2-set_2, CD4-extracel, DUF1968, Herpes_gE, Herpes_gI, Herpes_glycop_D, I-set, ICAM_N, ig, Ig_2, Ig_3, Ig_4, Ig_5, Ig_6, Ig_7, Ig_C17orf99, Ig_C19orf38, Ig_Tie2_1, Izumo-Ig, K1, Marek_A, ObR_Ig, PTCRA, Receptor_2B4, UL141, V-set, V-set_2, V-set_CD47

4.5.2 | Homologous immunoglobulin/non-immunoglobulin pairs

Homologous protein pairs in which either the query or the subject protein or both mapped to the immunoglobulin superfamily were classified as immunoglobulin (Ig) pairs. Protein pairs in which neither the query nor the subject protein mapped to this superfamily were classified as non-immunoglobulin (non-Ig) pairs (Figure 6, step 6).

4.6 | Calculation of difference distance maps (DDM) and DDM correlations

For each protein (PDBFlex cluster) with two conformations (subclusters), two distance maps (DMs) were calculated based on the representative coordinate sets of the two subclusters. These were then subtracted to get a difference distance map (DDM) that represented the conformational difference/change of the protein. Then, the similarity of the conformational changes of homologous proteins (i.e., different PDBFlex clusters) was assessed by calculating correlations between their DDMs (Figure 6, step 7). Several alignment steps and corrections were made to assure the proper assignment of equivalent residues in the four coordinate sets involved in each of these calculations. The technical description of these steps is given in the Supplementary Methods and in Figure S5 and Figure S6.

ACKNOWLEDGMENTS

We would like to thank Ms. Zhanwen Li for her assistance in utilizing the PDBFlex server for this project. We would also like to acknowledge our funding sources; NIGMS grant 118187 and the Bruce D. and Nancy B. Varner Endowment Fund.

AUTHOR CONTRIBUTIONS

Mallika Iyer: Conceptualization (lead); data curation (lead); formal analysis (lead); investigation (lead); methodology (lead); software (lead); validation (lead); visualization (lead); writing – original draft (lead); writing – review and editing (lead). **Lukasz Jaroszewski:** Conceptualization (equal); data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); resources (equal); software (equal); validation (equal); writing – original draft (equal). **Mayya Sedova:** Data curation (supporting); resources (supporting); software (supporting); validation (supporting). **Adam Godzik:** Conceptualization (equal); data curation (equal); formal analysis (equal); funding acquisition (lead); investigation (equal); methodology (equal); project administration (lead); resources (equal); supervision (lead); validation (equal); writing – original draft (equal); writing – review and editing (equal).

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

All output files, the script for creating/calculating subclusters for each PDBFlex cluster, and the script used for calculating DDMs and DDM correlations can be found at https://github.com/GodzikLab/ConformationalDiversity.

ORCID

Mallika Iyer https://orcid.org/0000-0003-0474-0594

REFERENCES

1. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J. 1986;5:823–826.
2. Brown KR, Jurisica I. Unequal evolutionary conservation of human protein interactions in interologous networks. Genome Biol. 2007;8:1–11.
3. Frauenfelder H, Sligar SG, Wolynes PG. The energy landscapes and motions of proteins. Science. 1991;254:1598–1603.
4. Cooper A. Protein fluctuations and the thermodynamic uncertainty principle. Prog Biophys Mol Biol. 1984;44:181–214.
5. Henzler-Wildman K, Kern D. Dynamic personalities of proteins. Nature. 2007;450:964–972.
6. Boehr DD, Nussinov R, Wright PE. The role of dynamic conformational ensembles in biomolecular recognition. Nat Chem Biol. 2009;5:789–796.
7. Eisenmesser EZ, Millet O, Labeikovsky W, et al. Intrinsic dynamics of an enzyme underlies catalysis. Nature. 2005;438:117–121.
8. Henzler-Wildman KA, Lei M, Thai V, Kerns SJ, Karplus M, Kern D. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. Nature. 2007;450:913–916.
9. Mulder FAA, Mittermaier A, Hon B, Dahlquist FW, Kay LE. Studying excited states of proteins by NMR spectroscopy. Nat Struct Biol. 2001;8:932–935.

10. Brändén G, Neutze R. Advances and challenges in time-resolved macromolecular crystallography. Science. 2021;373:eaba0954.
11. Rapaport DC. The art of molecular dynamics simulation. Cambridge, UK: Cambridge University Press, 2004.
12. Tirion MM. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys Rev Lett. 1996;77: 1905–1908.
13. Hinsen K. Analysis of domain motions by approximate normal mode calculations. Proteins Struct Funct Genet. 1998;33:417–429.
14. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J. 2001;80:505–515.
15. Kaynak BT, Zhang S, Bahar I, Doruker P. ClustENMD: Efficient sampling of biomolecular conformational space at atomic resolution. Bioinformatics. 2021;1–3:3956–3958. https://doi.org/10.1093/bioinformatics/btab496.
16. Fuglebakk E, Tiwari SP, Reuter N. Comparing the intrinsic dynamics of multiple protein structures using elastic network models. Biochim Biophys Acta. 2015;1850:911–922.
17. Ramanathan A, Agarwal PK. Evolutionarily conserved linkage between enzyme fold, flexibility, and catalysis. PLoS Biol. 2011; 9:e1001193.
18. Maguid S, Fernandez-Alberti S, Ferrelli L, Echave J. Exploring the common dynamics of homologous proteins. Application to the globin family. Biophys J. 2005;89:3–13.
19. Maguid S, Fernández-Alberti S, Parisi G, Echave J. Evolutionary conservation of protein backbone flexibility. J Mol Evol. 2006;63:448–457.
20. Zen A, Carnevale V, Lesk AM, Micheletti C. Correspondences between low-energy modes in enzymes: Dynamics-based alignment of enzymatic functional families. Protein Sci. 2008;17:918–929.
21. Gagné D, Charest LA, Morin S, Kovrigin EL, Doucet N. Conservation of flexible residue clusters among structural and functional enzyme homologues. J Biol Chem. 2012;287:44289–44300.
22. Maguid S, Fernandez-Alberti S, Echave J. Evolutionary conservation of protein vibrational dynamics. Gene. 2008;422:7–13.
23. Narunsky A, Nepomnyachiy S, Ashkenazy H, Kolodny R, Ben-Tal N. ConTemplate suggests possible alternative conformations for a query protein of known structure. Structure. 2015;23:2162–2170.
24. Sedova M, Jaroszewski L, Iyer M, Li Z, Godzik A. ModFlex: Towards function focused protein modeling. J Mol Biol. 2021; 433:166828.
25. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. Nucleic Acids Res. 2000;28:235–242.
26. Kosloff M, Kolodny R. Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins Struct Funct Genet. 2008;71: 891–902.
27. Burra PV, Zhang Y, Godzik A, Stec B. Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. Proc Natl Acad Sci. 2009;106:10505–10510.
28. Monzon AM, Zea DJ, Marino-Buslje C, Parisi G. Homology modeling in a dynamical world. Protein Sci. 2017;26:2195–2206.
29. Romesberg FE, Spiller B, Schultz PG, Stevens RC. Immunological origins of binding and catalysis in a Diels-Alderase antibody. Science. 1998;279:1929–1933.
30. Klarenbeek A, Mazouari KE, Desmyter A, et al. Camelid Ig V genes reveal significant human homology not seen in therapeutic target genes, providing for a powerful therapeutic antibody platform. MAbs. 2015;7:693–706.
31. Hrabe T, Li Z, Sedova M, Rotkiewicz P, Jaroszewski L, Godzik A. PDBFlex: Exploring flexibility in protein structures. Nucleic Acids Res. 2016;44:D423–D428.
32. Orellana L. Large-scale conformational changes and protein function: Breaking the in silico barrier. Front Mol Biosci. 2019; 6:117.
33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–410.
34. Mistry J, Chuguransky S, Williams L, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49:D412–D419.
35. www.hmmer.org.
36. Quiocho FA, Ledvina PS. Atomic structure and specificity of bacterial periplasmic receptors for active transport and chemotaxis: variation of common themes. Mol Microbiol. 1996;20: 17–25.
37. Fulyani F, Schuurman-Wolters GK, Žagar AV, Guskov A, Slotboom DJ, Poolman B. Functional diversity of tandem substrate-binding domains in ABC transporters from pathogenic bacteria. Structure. 2013;21:1879–1888.
38. Oh BH, Pandit J, Kang CH, et al. Three-dimensional structures of the periplasmic lysine/arginine/ornithine-binding protein with and without a ligand. J Biol Chem. 1993;268:11348–11355.
39. Oh BH, Ames GFL, Kim SH. Structural basis for multiple ligand specificity of the periplasmic lysine-, arginine-, ornithine-binding protein. J Biol Chem. 1994;269:26323–26330.
40. Petchey MR, Rowlinson B, Lloyd RC, Fairlamb IJS, Grogan G. Biocatalytic synthesis of Moclobemide using the amide bond Synthetase McbA coupled with an ATP recycling system. ACS Catal. 2020;10:4659–4663.
41. Reger AS, Wu R, Dunaway-Mariano D, Gulick AM. Structural characterization of a 140° domain movement in the two-step reaction catalyzed by 4-chlorobenzoate:CoA ligase. Biochemistry. 2008;47:8016–8025.
42. Gulick AM. Conformational dynamics in the acyl-CoA synthetases, adenylation domains of non-ribosomal peptide synthetases, and firefly luciferase. ACS Chem Biol. 2009;4:811–827.
43. Chen Q, Ji C, Song Y, et al. Discovery of McbB, an enzyme catalyzing the β-carboline skeleton construction in the marinacarboline biosynthetic pathway. Angew Chemie - Int Ed. 2013;52:9980–9984.
44. Stewart M. Structural basis for the nuclear protein import cycle. Biochem Soc Trans. 2006;34:701–704.
45. Forwood JK, Lonhienne TG, Marfori M, et al. Kap95p binding induces the switch loops of RanGDP to adopt the GTP-bound conformation: Implications for nuclear import complex assembly dynamics. J Mol Biol. 2008;383:772–782.
46. Liu SM, Stewart M. Structural basis for the high-affinity binding of nucleoporin Nup1p to the Saccharomyces cerevisiae importin-β homologue, Kap95p. J Mol Biol. 2005;349:515–525.
47. Tauchert MJ, Hémonnot, C, Neumann, P, Köster, S, Ralf Ficner, R, & Dickmanns, A. Impact of the crystallization condition on importin-β conformation: *Acta Crystallogr*. Sect D Struct Biol. 2016;72:705–717.
48. Li Z, Jaroszewski L, Iyer M, Sedova M, Godzik A. FATCAT 2.0: Towards a better understanding of the structural diversity of proteins. Nucleic Acids Res. 2020;48:60–64.

49. Weiss DR, Levitt M. Can morphing methods predict intermediate structures? J Mol Biol. 2009;385:665–674.
50. Park S, Schulten K. Calculating potentials of mean force from steered molecular dynamics simulations. J Chem Phys. 2004; 120:5946–5961.
51. Enosh A, Raveh B, Furman-Schueler O, Halperin D, Ben-Tal N. Generation, comparison, and merging of pathways between protein conformations: Gating in K-channels. Biophys J. 2008;95:3850–3860.
52. Daura, X., Gademann, K., Jaun, B., Gunsteren, W. F. Van & Mark, A. E. Peptide Folding: When Simulation Meets Experiment 38, 236–240 (1999).

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Iyer M, Jaroszewski L, Sedova M, Godzik A. What the protein data bank tells us about the evolutionary conservation of protein conformational diversity. Protein Science. 2022;31(7):e4325. https://doi.org/10.1002/pro.4325